Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«КУБАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»**

**(ФГБОУ ВО «КубГУ»)**

**Факультет компьютерных технологий и прикладной математики**

**Кафедра вычислительных технологий**

**ЛАБОРАТОРНАЯ РАБОТА №6.2**

**Дисциплина: Обработка больших данных**

Работу выполнил: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ А.М. Нагалевский

Направление подготовки: 02.03.02 Фундаментальная информатика и информационные технологии

Преподаватель: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В. И. Шиян

**Цель работы:** закрепить знания об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научиться визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты. Научиться выполнять классификацию на основе формулы Байеса и деревьев решений

**Вариант:** 18

**Часть 2.** **Байесовская классификация и деревья принятия решений на R**

Сначала построим таблицы вероятностей по всем характеристикам. В первом столбце – средние значения параметров, во втором столбце – их стандартные отклонения. Фрагмент таблицы можно увидеть на рисунке 15.

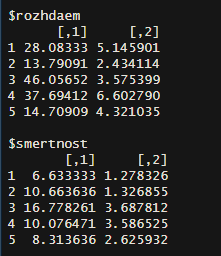
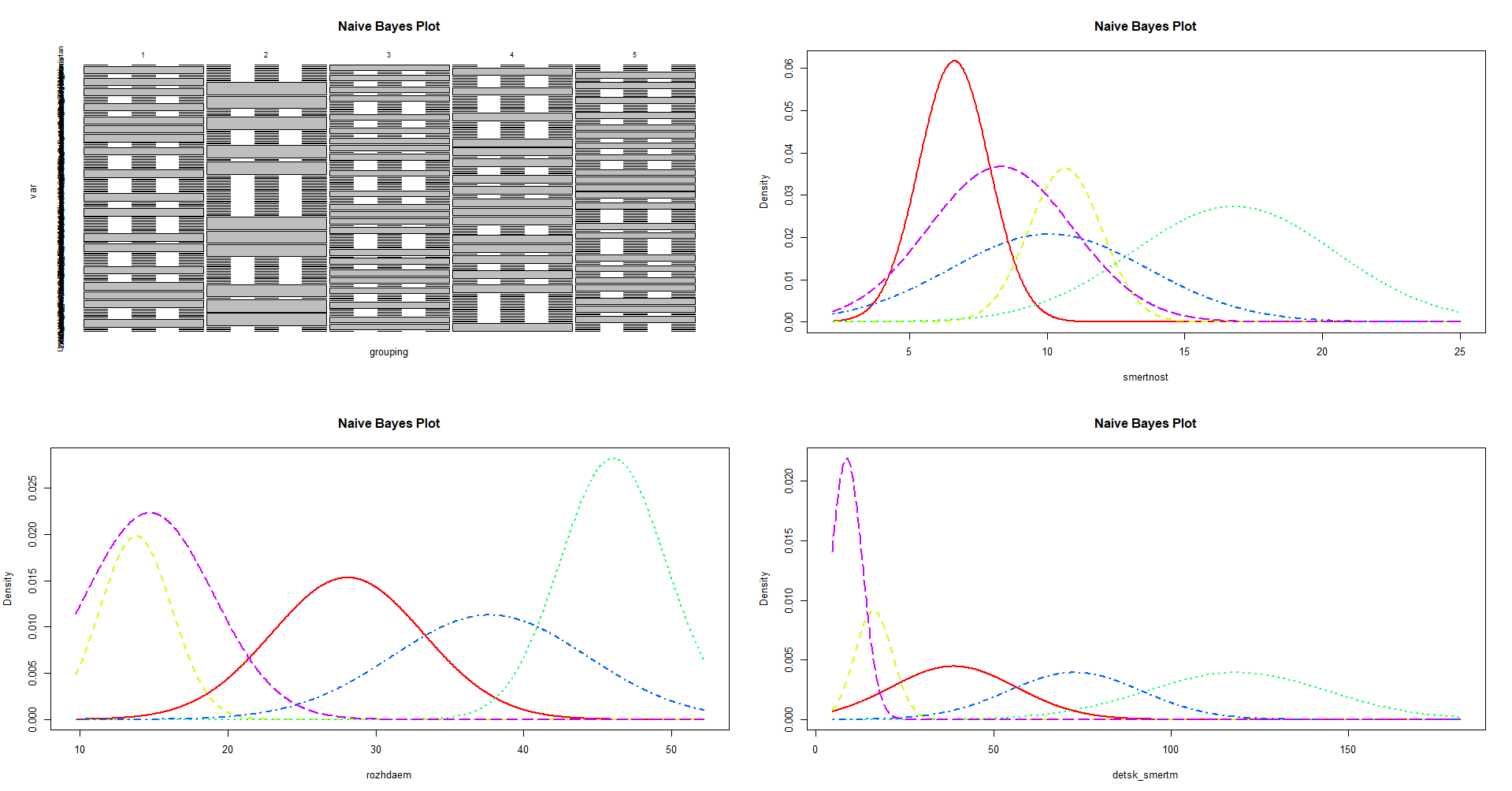


Рисунок 15 – Фрагмент таблицы вероятностей

Затем строятся функции плотности для характеристик на рисунке 16. Распределение на графике подтверждает правильность распределения на кластеры.



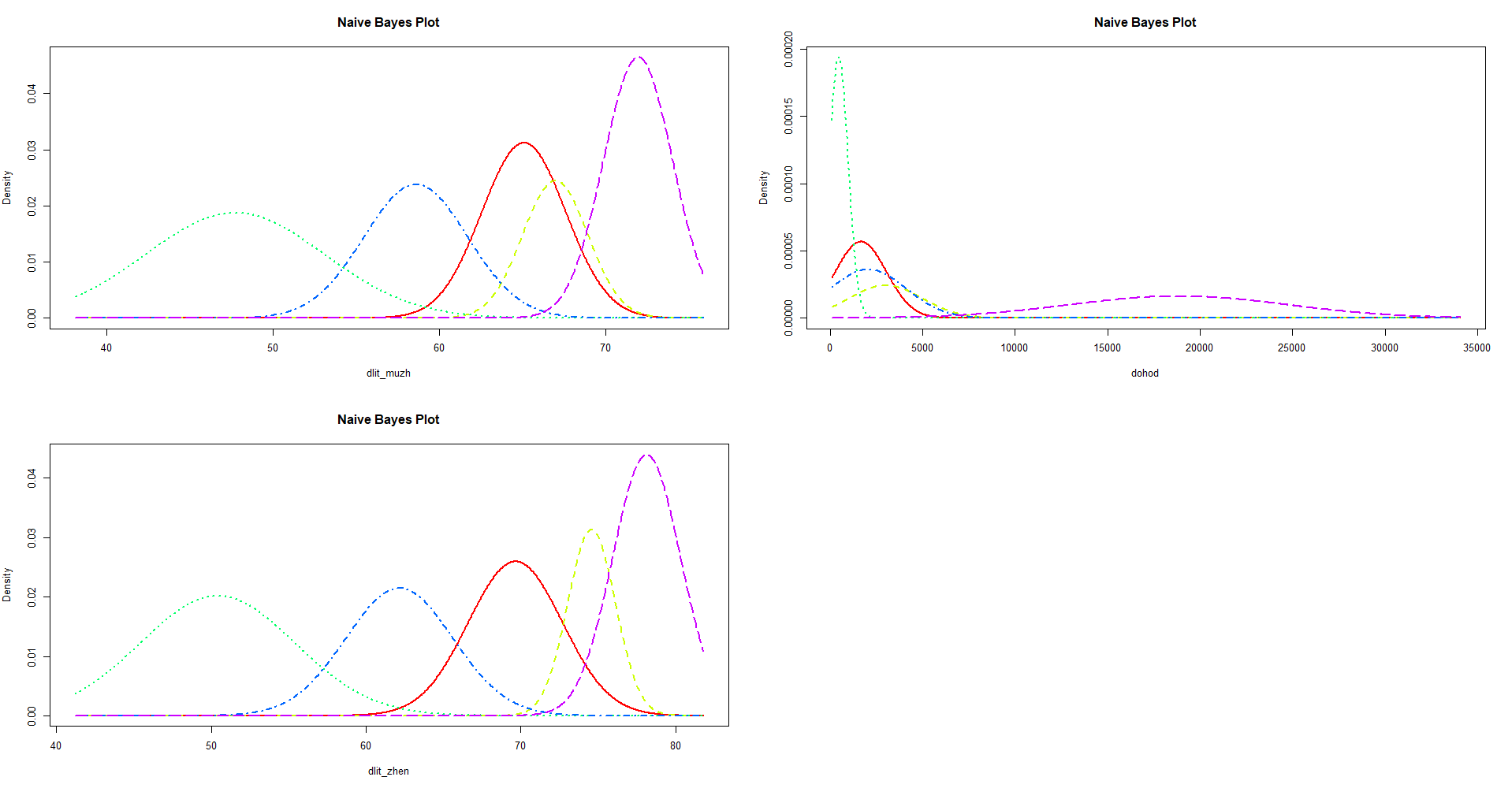


Рисунок 16 – Функции плотности

Далее проведем классификацию Decision Tree. Результаты можно увидеть на рисунке 17 и 18.

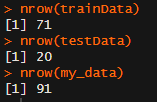


Рисунок 17 – Распределение филиалов

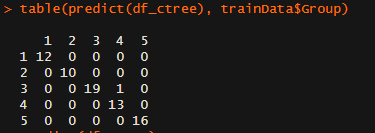


Рисунок 18 – Таблица с классификацией

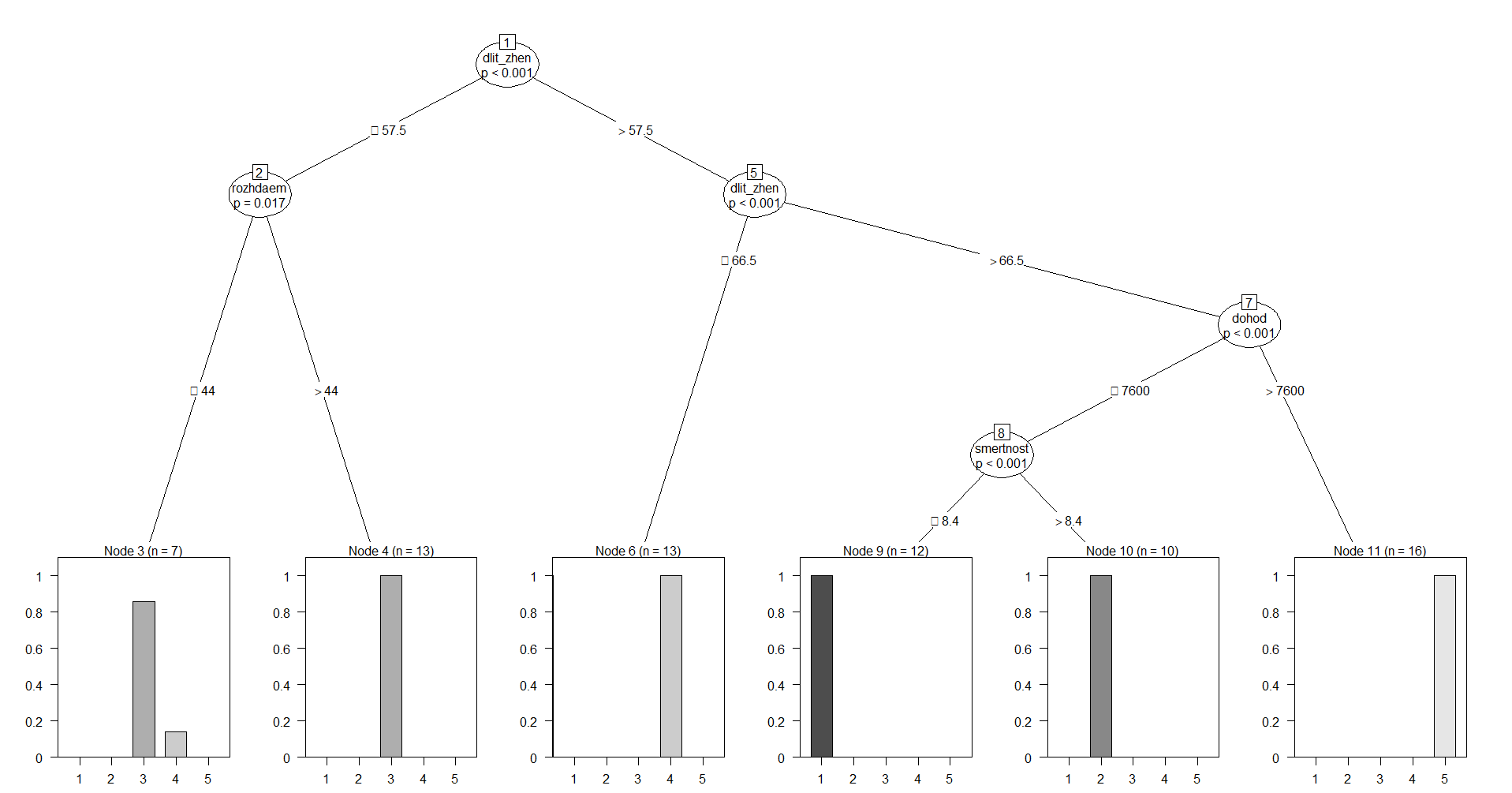


Рисунок 19- График полученный с помощью Decision Tree

Попробуем использовать более точный алгоритм Random Forest. На рисунке 20 изображена информация о модели, а именно об ошибках.

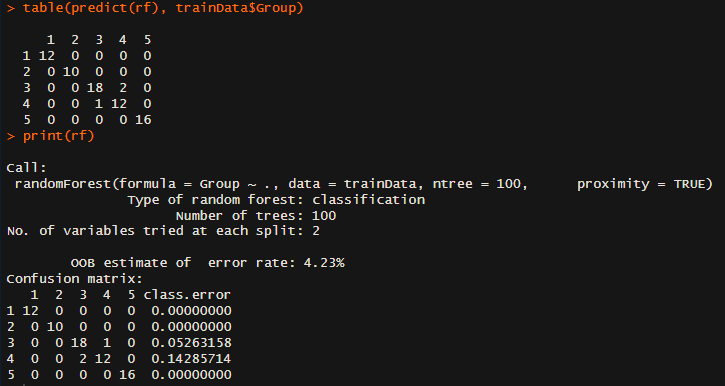


Рисунок 20– Результат работы алгоритма Random Forest

Алгоритм выделил 2 ошибки, процент ошибки 5.63%. При нескольких запусках процент ошибки вырастал максимум до 7%. Следовательно, предсказывать на таком количестве данных можно.

**Вывод**: были закреплены знания об алгоритмах классификации и кластеризации данных, ознакомиться с некоторыми функциями языка R, осуществляющими этот вид анализа, принципами их работы. Научился визуализировать результаты работы функций кластерного анализа и классификаторов, интерпретировать полученные результаты и выполнять классификацию на основе формулы Байеса и деревьев решений.